Computing Cost of Classification (Quiz)

+ : fraudulent transaction

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	x
	-	5	0

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

For what range of costs for False Negative is M1 better?

And for what range is M2 better?

Since M_2 makes more false negatives, if x was very large, we'd prefer M_1 (it'll have lower cost). However, there perhaps is a smaller value of x at which M_2 is as expensive as M_1. What is that value?

Cost of M_1: $-150 + 40x + 60 \times 5 + 250 \times 0 = 40x + 150$

Cost of M_2: $-250 + 45x + 5 \times 5 + 200 \times 0 = 45x - 225$

When the costs are equal

$$40x + 150 = 45x - 225 \Rightarrow 5x = 375 \Rightarrow x = 75$$

Thus, if

- $x < 75$ prefer M_2
- $x = 75$ indifferent between the two
- $x > 75$ prefer M_1

Tuning Prediction to Minimize Cost

- The cost of False Positive is \$12, cost of False Negative is \$3
  - If you are using a classifier that estimates the probability of a record being +ve, what is the lowest probability at which you'd classify a record to be +ve?

Hint: look for the probability threshold where the expected cost from mistake by either prediction is same. At that probability you'll be indifferent between predicting a record to be + or -

Record#	P(+)	Predict
1	0.99	+
2	0.91	+
3	0.84	?
4	0.75	?
	…	…
19	0.23	?
20	0.11	-

Let the threshold on the probability of a record being positive be p. At this probability the expected cost from predicting positive and from predicting negative are the same.

The mistake we can make from predicting positive is if the record is negative, I.e., we'd be making a false positive mistake. The probability of this mistake is $1 - p$ and the cost is \$12.

Similarly, the mistake we can make from predicting negative is a false negative at a probability of p (which is the probability of the record being positive, against our prediction) and cost \$3.

The threshold can be calculated by equating the expected costs and solving for p.

$$(1 - p) \times C_{FP} = p \times C_{FN}$$

$$\Rightarrow C_{FP} = (C_{FP} + C_{FN}) \times p \Rightarrow p = \frac{C_{FP}}{C_{FP} + C_{FN}} = \frac{12}{15} = 0.8$$

We'd be minimizing our expected costs from wrong predictions if we predict positive for all records with probability of being positive greater than 0.8 and predict negative for all records with probability of being positive less than 0.8.